2.1.5. Histogramas

2.1.5.1. Introdução

Entre os gráficos discutidos nesse capítulo, o *histograma* (*histogram*) é o mais tradicional, freqüentemente usado para mostrar, por exemplo, dados de recenseamentos ou de saúde pública.

O histograma é a representação gráfica de uma *distribuição de freqüências* de dados agrupados. Nele, o eixo horizontal é dividido em subintervalos correspondentes às classes, e a freqüência de cada classe é representada pelas áreas dos retângulos construídos sobre estes subintervalos. Note que o importante é a *área* do retângulo, não a *altura*. É claro que, se os dados estiverem organizados em classes de intervalos iguais, as bases dos retângulos serão iguais, e as áreas serão proporcionais às alturas; neste caso, as representações das freqüências por áreas ou por alturas serão equivalentes. (Esta aliás é uma das razões pelas quais é preferível que as classes de uma tabela de distribuição de frequências tenham o mesmo intervalo: fica fácil ao leitor imaginar a forma do histograma ao examinar a tabela).

Um mesmo histograma pode representar tanto a freqüência *absoluta* quanto a freqüência *relativa* ou a *percentual*, já que estas freqüências são proporcionais entre si. As formas do histograma que representam estes três tipos de freqüências serão portanto idênticas, sendo mudada apenas a escala do eixo vertical.

2.1.5.2. Como fazer o histograma

O único problema, se quisermos desenhar um histograma, será o de decidir quantas classes devem ser usadas. Na Seção **2.1.4.4**, observamos que, para uma tabela de distribuição de frequências, não existe uma regra fixa para determinar o número de classes (NC) em que o domínio da variável deve ser dividido (ou, equivalentemente, o número de retângulos do histograma), e que isto depende da quantidade de dados da distribuição, e dos objetivos dos pesquisadores.

Como exemplo, o diagrama de pontos da Fig. **1** mostra os pesos de 493 recém-nascidos. Poderíamos representar estes dados numa distribuição de freqüências com 50 classes, cada qual com intervalo de classe (IC) de aproximadamente 100 gramas. O resultado seria um histograma como o da Fig. **2A**. Este histograma certamente representa de modo muito fiel a distribuição da amostra original (seu perfil é praticamente idêntico ao do gráfico de pontos na Fig. **1**).

Contudo, esta representação provavelmente não seria muito útil, por duas razões. Primeiro, por uma razão prática: a tabela de frequências em que este histograma se baseia é extensa demais - tem 50 linhas, e não caberia numa página impressa comum. Segundo, por

uma razão teórica: na maioria das vezes, analisamos uma *amostra* para conhecer a *população* de onde ela foi retirada. Os detalhes da amostra são acidentais, devidos ao acaso, e geralmente não interessam muito; se retirarmos uma outra amostra, provavelmente serão um pouco diferentes.

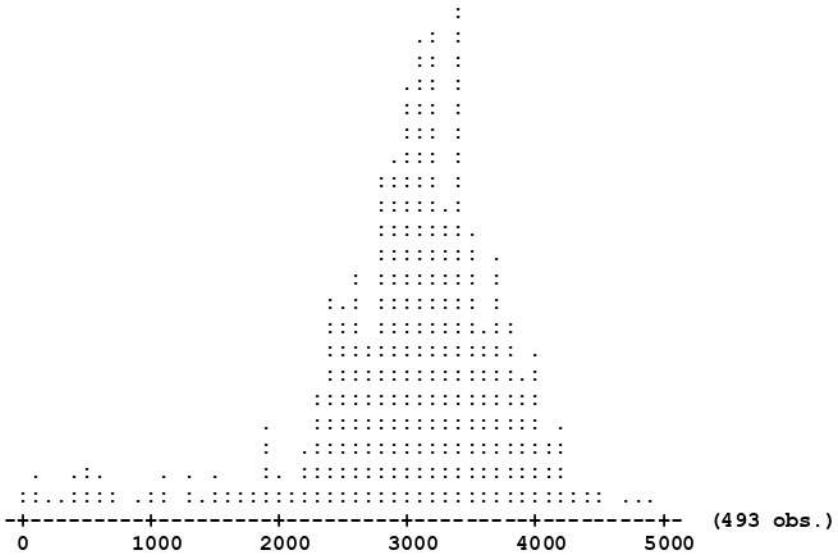
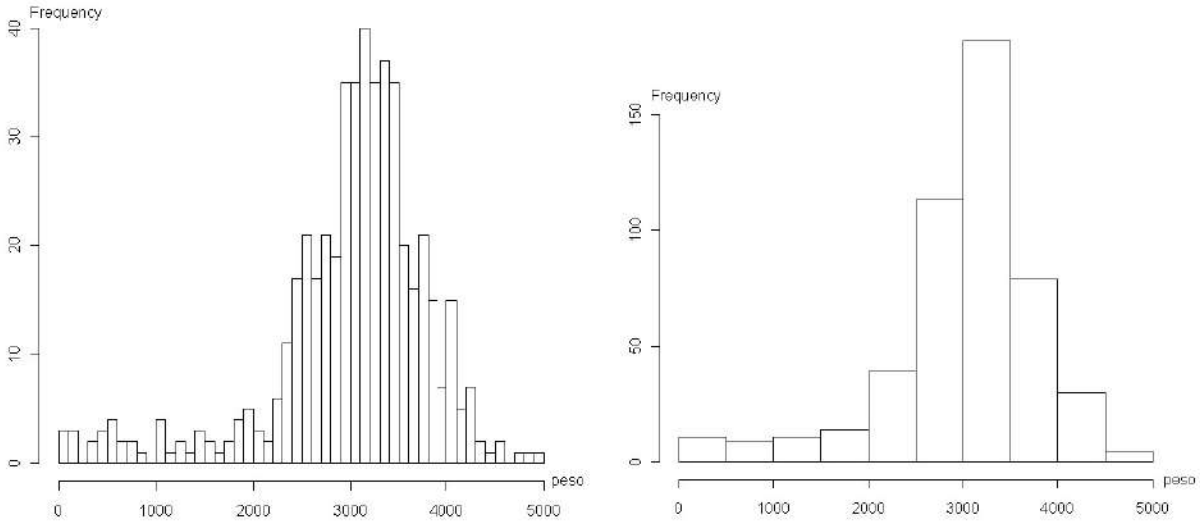


Figura 1. Peso ao nascer de crianças (g)

Neste gráfico, por exemplo, existe um "vale" cercado por dois picos, entre 3300 e 3400 g; este é um resultado acidental desta amostra, que quase certamente não reflete o que acontece na população (é pouco provável que haja, na população, muita diferença entre o número de crianças que nascem com 3350 g, e o de crianças que nascem com 3450 g!).



A - Histograma com 50 classes **B - Histograma com 10 classes**

Figura 2. Peso ao nascer de 493 crianças (g)

O que nos interessa é obter uma indicação aproximada de qual deve ser a forma geral da distribuição da população (ou seja, dos pesos de *todas* as crianças recém-nascidas, não apenas das crianças desta amostra), para que possamos encontrar um modelo probabi-

lístico para esta distribuição, e usá-lo na *Inferência Estatística* (Cap. **4**). Não precisamos portanto de um gráfico muito complicado - o excesso de detalhes só vai atrapalhar, pois irá obscurecer a forma básica. Lembre-se de que, sempre que fazemos uma análise estatística de dados, estamos querendo *mostrar* algo; o melhor gráfico é aquele que deixa mais evidente o que queremos mostrar; não necessariamente o mais complicado.

No exemplo, um histograma com menor número de classes provavelmente seria mais útil; a Fig. **2B** mostra novamente os dados, desta vez organizados em 10 classes, de intervalo 500 g cada. A forma geral da distribuição é a mesma da do gráfico anterior (unimodal, com leve assimetria negativa), mas está mostrada de modo muito mais claro. A tabela de frequências, por outro lado, seria menor (10 linhas), e não haveria problemas de espaço na página.

2.1.5.3. Relação entre o histograma e os gráficos de pontos e de ramo-e-folhas

Tanto o histograma quanto o diagrama de ramo-e-folhas representam os dados de forma *agrupada*; o gráfico de pontos também pode ser usado às vezes desta forma. Os três gráficos podem dar resultados parecidos. Por exemplo, voltemos aos dados da Amostra A, mostrada na Seção **2.1.3.2**, e o diagrama de ramo-e-folhas correspondente, reproduzido na Fig. **3**.

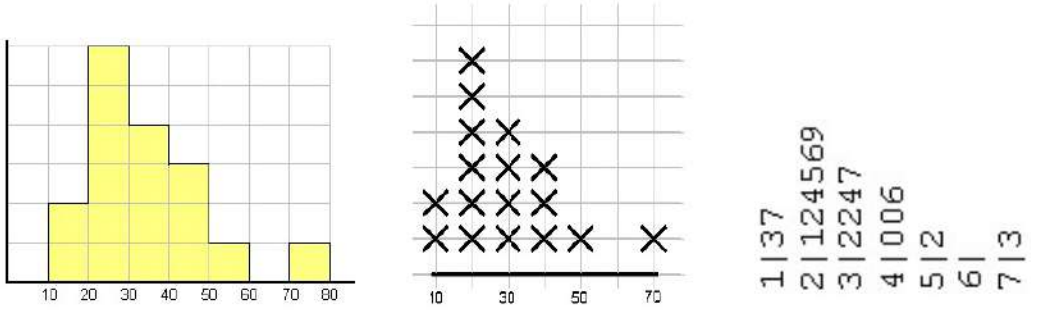
```
Amostra A: 13 17 21 22 24 25 26 29 32 32 34 37 40 40 46 52 73

1|37
2|124569
3|2247
4|006
5|2
6|
7|3
```

Figura 3. Diagrama de ramo-e-folhas das idades (em anos) dos pacientes, Amostra A

Se agruparmos os dados usados em classes de intervalo IC =10, obteremos a Tab. **1**. Na Fig. **4A** está o histograma que representa esta tabela; na Fig. **4B**, um gráfico de pontos, onde os valores foram truncados para a dezena mais abaixo (o que é equivalente ao que foi feito na tabela); na Fig. **4C**, o diagrama de ramo-e-folhas da Fig. **3**, mas agora com o eixo na horizontal. Os três gráficos tem a mesmo contorno, e transmitem a mesma informação.

Tabela 1. Idades dos pacientes Amostra A

idade	f
10-19	2
20-29	6
30-39	4
40-49	3
50-59	1
60-69	-
70-79	1
	17



A. Histograma **B. Gráfico de pontos** **C. Gráfico de ramo-e-folhas**

Figura 4. Comparação dos três gráficos

O histograma é mais flexível, uma vez que não é obrigatório que os dados sejam agrupados em dezenas (ou centenas) ou seus submúltiplos; poderíamos, por exemplo, agrupar os dados em classes com intervalo igual a 4, como em 12-16, 16-20, 20-24, etc. (o que pode ser útil às vezes, mas em geral **não** é uma boa idéia, pois a divisão em múltiplos de 5 ou 10 parece sempre mais natural). Além disso, o histograma pode ser usado para representar grandes quantidades de dados (o gráfico de pontos ou o de ramo-e-folhas ficariam sobrecarregados, se a amostra tivesse milhares de valores).

Por outro lado, o diagrama de ramo-e-folhas tem a vantagem de permitir que os valores originais dos dados sejam recuperados a partir do gráfico, o que não é possível no histograma ou no gráfico de pontos. Nas Figs. **4A** e **4B**, por exemplo, vemos que há duas observações na primeira classe, mas não sabemos quais são seus valores exatos; no ramo-e-folhas da Fig. **4C**, vemos imediatamente que os valores são 13 e 17.

2.1.5.4. Histogramas com classes de intervalos diferentes

Se os intervalos de classes forem diferentes, as áreas dos retângulos não serão mais proporcionais às alturas. As *áreas* representarão as *frequências* das classes; as *alturas* representarão as *densidades de freqüência* das classes. Como num retângulo temos que

área = altura x base
altura = área / base

e no histograma

área → freqüência (f)
base → intervalo de classe (IC)

obtemos a relação

densidade = freqüência / intervalo de classe = f / IC

No histograma da Fig. **5**, por exemplo, os retângulos *C* e *E* representam classes de mesma freqüência ($f = 6$) e têm portanto a mesma área.

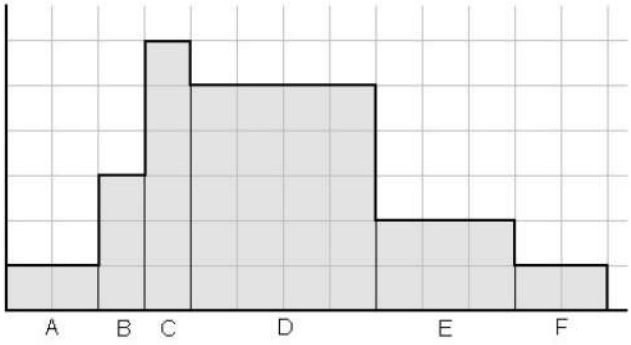


Figura 5. Histograma com classes de intervalos desiguais

Como o retângulo *C* tem base menor que o retângulo *E*, sua altura terá de ser maior; dizemos então que a classe representada por *C* tem maior "densidade" do que a representada por *E*, porque nela a mesma quantidade de dados se encontra concentrada num intervalo menor (o que deixa a distribuição mais "densa" neste intervalo). Este conceito de "densidade" é bastante importante, e voltaremos a ele quando discutirmos a *densidade de probabilidade* de variáveis aleatórias (seção **3.4.1.2**)

A representação por ICs diferentes tem a desvantagem de ser mais trabalhosa. Primeiro, porque a escolha dos ICs e dos limites das classes tem que ser feita pelos pesquisadores, e não automaticamente por um computador. Segundo, porque torna a interpretação da tabela mais difícil. Não é possível, se examinamos apenas as frequências numa tabela, saber quais seriam as classes mais *densas*, correspondentes a retângulos de maiores alturas. No exemplo, as classes *C* e *E* têm freqüências iguais, mas são representadas no histograma por retângulos de formas diferentes. Mesmo assim, este tipo de gráfico é útil, e às vezes pode ser a única maneira de representar convenientemente os dados.

Tabela 2 - Nascidos vivos em Minas Gerais, 2018

Peso ao nascer (g)	f	fr	fr(%)	F	F%
0 |-- 500	378	0,0014	0,14	378	-
500 |-- 1000	1.494	0,0057	0,57	1.872	0,71
1000 |-- 1500	2.130	0,0081	0,81	4.002	1,52
1500 |-- 2500	20.727	0,0786	7,86	24.729	9,40
2500 |-- 3000	66.022	0,2504	25,04	90.751	34,42
3000 |-- 4000	163.050	0,6185	61,85	253.801	96,27
4000 ou mais	9.831	0,0373	3,73	263.632	100,00
Ignorado	8	0,0000	0,00	263.640	100,00
Totais	263.640	1,0000	100,0	---	---

(fonte: datasus.gov.br)

Um caso particular de distribuições com ICs desiguais são as dos pesos ao nascer de crianças, como publicadas pelo SUS. A Tabela **2** (já mostrada na seção **2.1.4.3**) mostra os pesos de todas as crianças nascidas no Brasil em 2018; a Tab. **3**, os das crianças nascidas em Juiz de Fora (MG) em 2001.

Ambas as distribuições têm também classes "abertas" em seus extremos, isto é, classes que não tem definidos o limite inferior ou o superior. Apesar de classes abertas serem muito encontradas na prática, é melhor evitar usá-las; uma vez que não há limites nas classes extremas, não podemos construir o histograma (as dimensões das bases dos retângulos nas extremidades serão desconhecidas). Além disso, não podemos estimar a média da distribuição a partir da tabela (Seção **2.2.1.5**). Outro detalhe importante destas

tabelas é terem classes com crianças cujo peso é "ignorado"; estas classes não serão representadas no histograma, e também não serão levadas em conta no cálculo da freqüência relativa *fr*.

Tabela 3. Nascidos vivos em Juiz de Fora, 2001

Peso ao nascer (g)	f	fr	fr(%)	F	F%
0 ⊢ 500	6	0,0006	0,06	6	0,06
500 ⊢ 1000	60	0,0062	0,62	66	0,68
1000 ⊢ 1500	90	0,0092	0,92	156	1,60
1500 ⊢ 2500	974	0,1001	10,01	1130	11,61
2500 ⊢ 3000	2744	0,2820	28,20	3874	39,81
3000 ⊢ 4000	5562	0,5715	57,15	9436	96,96
4000 ou mais	296	0,0304	3,04	9732	100,00
Ignorado	31	-		9763	-
Totais	9763	1,0000	100,0	-	-

(fonte: datasus.gov.br)

2.1.5.5. Polígono de freqüências

Outra forma de representar graficamente uma distribuição de freqüências é através do *polígono de freqüências* (Fig. **6B**). Este gráfico é construído a partir do histograma, bastando que se liguem os pontos médios do topo dos retângulos consecutivos; os pontos médios do primeiro e do último retângulos são ligados a pontos no eixo horizontal que corresponderiam aos pontos médios de classes fictícias localizadas antes da primeira e depois da última classe da distribuição, respectivamente.

Note que a área do polígono de freqüências é idêntica a do histograma; examinando a Fig. **6A**, podemos notar que os trechos em que a área do histograma excede a do polígono correspondem exatamente a trechos em que a área do polígono excede a do histograma. Como o histograma, o polígono de frequências também pode ser usado para representar tanto a freqüência absoluta quanto a freqüência relativa ou a percentual; a interpretação de ambos os gráficos é a mesma.

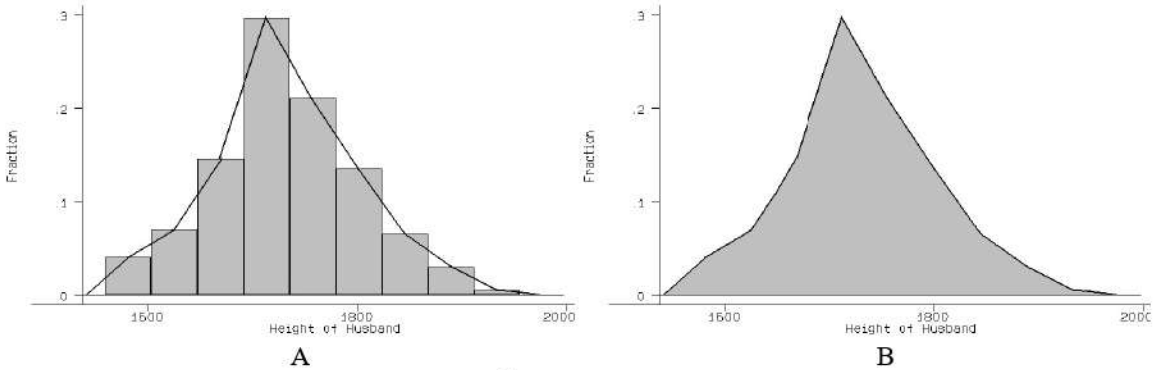


Figura 6. Polígono de freqüências – Construção a partir do histograma

O histograma é mais comumente usado, e é mais conhecido do público; o polígono, porém, tem a vantagem de permitir a comparação de duas distribuições, por meio de gráficos superpostos. A Fig. **7**, por exemplo, compara as alturas de homens e mulheres, numa amostra de casais ingleses; as duas distribuições tem formas similares, mas posições distintas (os homens são em média cerca de 10 cm mais altos do que as mulheres).

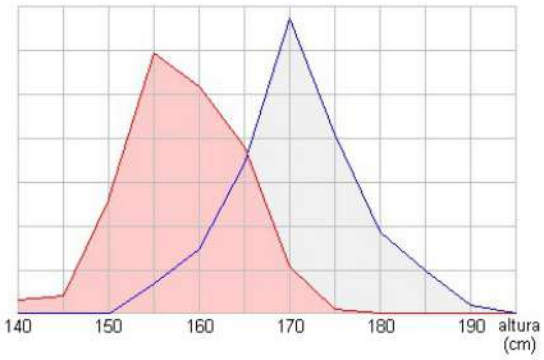


Figura 7. Alturas : homens (azul) x mulheres (vermelho)

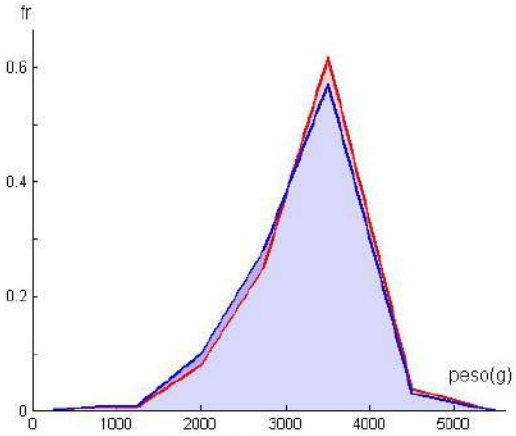
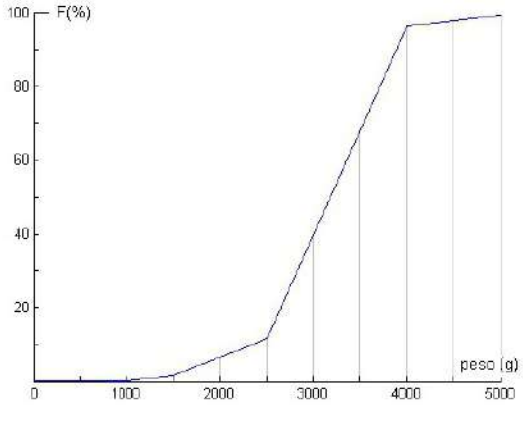


Figura 8. Peso ao nascer
J. Fora, 2001 (azul) x Brasil, 2018 (vermelho)

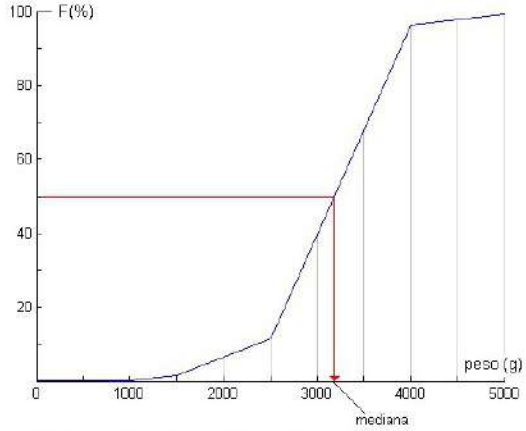
A Fig. **8** compara os pesos ao nascer de crianças nascidas no Brasil em 2018 (Tab. **2**) com os das crianças nascidas em Juiz de Fora em 2001 (Tab. **3**). Apesar de uma tabela se referir ao país todo, e a outra a apenas uma cidade, e de representarem épocas muito diferentes (2001 e 2018), ambas as distribuições tem formas praticamente idênticas.

2.1.5.6. Ogiva de Galton

As tabelas de distribuição de freqüências podem incluir um tipo de freqüência cuja representação ainda não foi apresentada: as *freqüências acumuladas* absolutas, relativas ou percentuais. O gráfico usado para representá-las é o *polígono de freqüências acumuladas* ou *ogiva de Galton*.



A. Ogiva de Galton – peso ao nascer



B. Localização gráfica da mediana na ogiva

Figura 9. Ogiva de Galton

A Fig. **9** mostra a ogiva para os dados de peso ao nascer da Tab. **3**. Neste gráfico, são marcados os pontos definidos pelos limites superiores de cada classe no eixo horizontal, e as freqüência acumulada destas classes no eixo vertical; o gráfico é feito unindo-se estes pontos por uma linha poligonal. As ogivas contém a mesma informação que o histograma ou o polígono de freqüências (já que as freqüências acumuladas podem ser calculadas a partir das freqüências absolutas, ou vice-versa), mas esta informação é representada de forma totalmente diversa.

A principal utilidade das ogivas é permitir a localização rápida das *separatrizes*. Por exemplo, se quisermos determinar aproximadamente qual deve ser a mediana da distribuição, basta procurar o valor da variável relacionado a uma freqüência acumulada percentual de 50%; no gráfico (Fig. **9B**), vemos que é de aproximadamente 3200 g; isto quer dizer que metade destas crianças nasceram com pesos inferiores a 3200 g.

2.1.5.7. Pirâmides etárias

A *pirâmide etária* é um gráfico que apresenta a distribuição das idades dos habitantes de um país ou região. São feitos dois histogramas separados, um com as idades dos homens, agrupadas em classes de intervalos de 5 anos, e outro com as idades das mulheres. Em seguida, estes dois histogramas são justapostos, com os eixos que representam as *idades* colocados na vertical, e os eixos que representam as *densidades de freqüência* colocados na horizontal, em sentidos opostos.

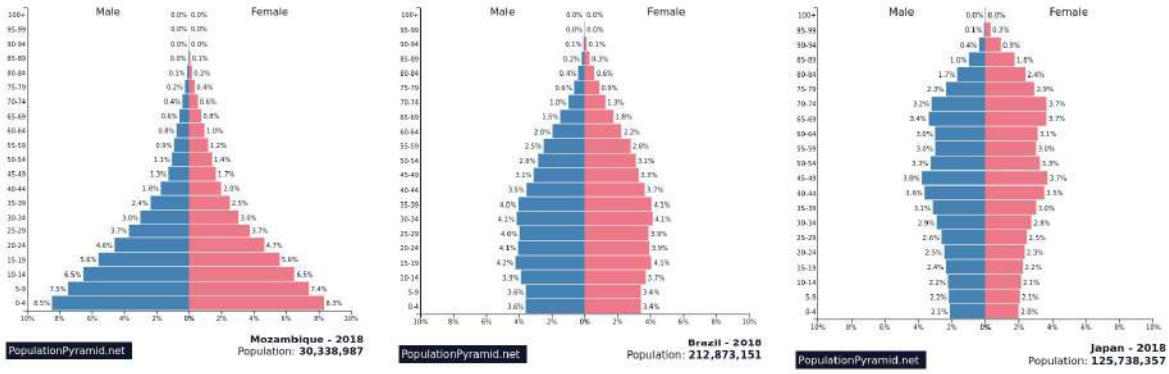


Figura 10. Pirâmides etárias de Moçambique, Brasil e Japão
(fonte: *PopulationPyramid.net*)

A Fig. **10** mostra as pirâmides etárias de Moçambique, Brasil e Japão. O nome "pirâmide etária" se justifica no caso do gráfico que representa Moçambique. Contudo, a medida que um país se desenvolve, a expectativa de vida de seus habitantes tende a crescer, enquanto a taxa de natalidade tende a diminuir; o resultado é que o gráfico perde a sua forma triangular, e as frequências maiores são encontradas em idades cada vez mais avançadas, a medida que a população envelhece. Os gráficos mostram claramente que o Brasil fica numa posição intermediária entre Moçambique (um país de população jovem, e expectativa de vida comparativamente baixa) e o Japão (a terceira maior economia, e a maior expectativa de vida do mundo).

Resumo

- O histograma é a representação gráfica da informação contida numa *tabela de distribuição de frequências*. A freqüência de cada classe é representada pela *área* do retângulo construído sobre o intervalo desta classe, não pela *altura* do retângulo.
- O histograma mostra detalhes sobre a *forma* de uma distribuição: indica se há mais de um aglomerado, se a distribuição é simétrica ou não, unimodal ou não, etc.
- Deve ser usado para representar grandes quantidades de dados (de populações, ou de amostras muito grandes); não vale a pena usá-lo para representar amostras pequenas.
- Para comparar duas distribuições, pode ser usado o *polígono de frequências*, derivado do histograma.
- O histograma é o mais tradicional dos gráficos que mostram a forma de uma distribuição, e é por isso bem conhecido do público (ao contrário dos *diagramas de Tukey* e *ramo-e-folhas*).